Using parallel treebanks for machine translation evaluation

Plamada, Magdalena ; Volk, Martin

Abstract: This paper presents a new method to evaluate machine translation (MT) systems against a parallel treebank. This approach examines specific linguistic phenomena rather than the overall performance of the system. We show that the evaluation accuracy can be increased by using word alignments extracted from a parallel treebank. We compare the performance of our statistical MT system with two other competitive systems with respect to a set of problematic linguistic structures for translation between German and French.

Using Parallel Treebanks for Machine Translation Evaluation

Magdalena Plamadă, Martin Volk

Institute of Computational Linguistics
University of Zurich
{plamada, volk}@cl.uzh.ch

Abstract

This paper presents a new method to evaluate machine translation (MT) systems against a parallel treebank. This approach examines specific linguistic phenomena rather than the overall performance of the system. We show that the evaluation accuracy can be increased by using word alignments extracted from a parallel treebank. We compare the performance of our statistical MT system with two other competitive systems with respect to a set of problematic linguistic structures for translation between German and French.

1 Introduction

An important step in improving the performance of a statistical machine translation (SMT) system is the diagnosis of its output. As human evaluation is expensive and the automatic metrics fail to convey information about the nature of the errors, researchers in the field have worked on linguistically-informed evaluation measures. The advantage of this approach is that it can pinpoint the weaknesses of MT systems in terms of morpho-syntactic errors.

Liu and Gildea [7] were among the first to incorporate syntactic features (and dependency relations) into MT evaluation. Their approach correlated better with the human judgments than the previous n-gram based metrics (e. g. BLEU) and has thus underlain following research in the field. Furthermore, semantic-based evaluation metrics such as [8] were developed with the purpose of assessing the meaning similarity. Latest approaches describe an evaluation metric which aims at incorporating several levels of linguistic information (lexical, morphological, syntactical and semantical) [3].

Although these metrics reflect various linguistic levels, they cannot perform a real diagnosis of MT systems. We therefore need a thorough analysis focused on different linguistic levels. In this paper, however, we only refer to the diagnosis of morpho-syntactic errors. Popović and Ney [11] proposed a method for identifying and analyzing translation errors involving different Part-of-Speech (PoS) classes.

Zhou et al. [15] introduced the idea of diagnostic evaluation based on linguistic checkpoints (see section 3) and released it as a stand-alone tool: Woodpecker[1]. Unfortunately, their tool works only for English-Chinese and is released under a restrictive license. On the other hand, a freely-available software, DELiC4MT[2], offers the same functionalities plus the option of adapting it to any language pair.

This paper builds upon previous research on linguistic checkpoints. Since this type of evaluation involves a fine-grained analysis of the texts in the source and target language, word correspondence is a very important prerequisite. Moreover, the quality of the evaluation strongly depends on the accuracy of these alignments. As both approaches use automatic alignment methods, the accuracy of the resulting alignments decreases. Therefore we suggest to avoid this drawback by extracting good alignments from a manually-checked parallel treebank.

This paper is structured as follows: In section 2 we describe our data and in the subsequent one the evaluation process. Section 4 introduces the changes we have made to the existing evaluation workflow. Section 5 presents and analyzes our experimental efforts. Finally, section 6 wraps up the discussion.

2 Our Reference Corpus: The Alpine Parallel Treebank

The reported experiments have been carried out on the German-French parallel treebank part of the SMULTRON corpus[3]. The treebank consists of 1000 sentences from the Text+Berg corpus[4], which contains the digitized publications of the Swiss Alpine Club from 1864 until 2011. The parallel treebank contains the "same" text in German and French, with most texts being translated from German into French and only a few of them vice versa.

We refer to a treebank as to a particular kind of annotated corpus where each sentence is mapped to a graph (a tree) which represents its syntactic structure. In addition to the syntactic annotation, the parallel treebank is aligned on the sub-sentential level, for example on the word or the phrase level. We regard phrase alignment as alignment between linguistically motivated phrases and not just arbitrary consecutive word sequences, as in statistical machine translation.

The annotation is a semi-automatic process, as we have manually checked and corrected the annotations at each processing step. PoS tagging is performed by the TreeTagger[5], with its standard parameter files for German and our in-house trained parameters for French, respectively. The tagged texts are then loaded into Annotate[6], a treebank editor which suggests constituent phrases and function labels based, in German, on the structures provided by the TnT Chunker[7]. For French, the

[1] http://research.microsoft.com/en-us/downloads/ad240799-a9a7-4a14-a556-d6a7c7919b4a
[2] http://www.computing.dcu.ie/~atoral/delic4mt/
[3] http://www.cl.uzh.ch/research/paralleltreebanks/smultron_en.html
[4] http://www.textberg.ch
[5] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger
[6] http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/annotate.html
[7] http://www.coli.uni-saarland.de/~thorsten/tnt/

phrases are generated by a shallow parsing model (hidden Markov model) trained on the Le Monde corpus [1]. Finally, the monolingual treebanks are exported in the TIGER-XML format [6]. More details about the annotation of the German treebank can be found in [14], whereas the French annotation is described in [5].

We use the Bleualign algorithm [13] to align the sentences across both monolingual treebanks. Our alignment convention was to discard the automatic many-to-many word alignments for the purpose of increasing the precision. Subsequently, a human annotator checked and, when required, corrected the remaining word and sentence alignments and then added the phrase alignments. Finally, the alignment file is available in XML format, as the following snippet shows:

```
<align type="good" last_change="2010-09-03">
      <node treebank_id="de" node_id="s225_18"/>
      <node treebank_id="fr" node_id="s231_16"/>
</align>
```

This says that node 18 in sentence 225 of the German treebank (de) is aligned with node 16 in sentence 231 of the French treebank (fr). The node identifiers refer to the IDs in the TIGER-XML treebanks. The alignment is labeled as *good* when the linked phrases represent exact translations and as *fuzzy* in case of approximate correspondences.

3 The Evaluation Tool: DELiC4MT

DELiC4MT is an open-source tool that performs diagnostic evaluation of MT systems over user-defined linguistically-motivated constructions, also called *checkpoints*. This term was introduced by Zhou et al. [15] and refers to either lexical elements or grammatical constructions, such as ambiguous words, noun phrases, verb-object collocations etc. The experiments reported in this paper follow the workflow proposed by Naskar et al. [9], due to its option to integrate new language pairs.

First the texts are PoS-tagged and exported in the KYOTO Annotation Format (KAF)[2]. This scheme facilitates the inspection of the terms in the sentences and thus querying for specific features, such as PoS sequences. Figure 1 depicts the KAF annotation of the German phrase *den ersten Gipfel* (EN: the first peak) and its French equivalent *le premier sommet*.

The linguistic checkpoints are subsequently defined in the so-called kybot profiles. A kybot profile starts with the declaration of the involved variables and the relations among them and ends specifying which attributes of the matched terms should be exported. For example, figure 2 depicts the kybot profile for a nominal group consisting of a determiner, an adjective and a noun. Moreover, the constituent terms have to be consecutive. Once defined, the kybot profile is run over the source language KAF files and the matched terms (with the specified attributes) are stored in a separate XML file.

```
<text>[...]
  <wf wid="w729_6" sent="729" para="1">den</wf>
  <wf wid="w729_7" sent="729" para="1">ersten</wf>
  <wf wid="w729_8" sent="729" para="1">Gipfel</wf>
[...]</text>
<terms>[...]
 <term tid="t729_6" type="open" lemma="d" pos="ART">
  <span> <target id="w729_6"/> </span>
 </term>
 <term tid="t729_7" type="open" lemma="erst" pos="ADJA">
  <span> <target id="w729_7"/> </span>
 </term>
 <term tid="t729_8" type="open" lemma="Gipfel" pos="NN">
  <span> <target id="w729_8"/> </span>
 </term>
[...]</terms>

<text>[...]
  <wf wid="w729_6" sent="729" para="1">le</wf>
  <wf wid="w729_7" sent="729" para="1">premier</wf>
  <wf wid="w729_8" sent="729" para="1">sommet</wf>
[...]</text>
<terms>[...]
 <term tid="t729_6" type="open" lemma="le" pos="DET:ART">
  <span> <target id="w729_6"/> </span>
 </term>
 <term tid="t729_7" type="open" lemma="premier" pos="NUM">
  <span> <target id="w729_7"/> </span>
 </term>
 <term tid="t729_8" type="open" lemma="sommet" pos="NOM">
  <span> <target id="w729_8"/> </span>
 </term>
[...]</terms>
```

Figure 1: Sample KAF annotation for a German-French sentence pair

The last step evaluates how well did the MT system translate the linguistic phenomena of interest. The evaluation is based on n-gram similarity, thus counting the overlapping word sequences between the hypothesis (automatic translation) and the reference (the previously identified checkpoint instances). The evaluation module requires as input the source and target language texts in KAF format, as well as the word alignments between them, the XML file produced at the previous step and the automatic translation to be evaluated. Each matched instance is evaluated separately and, on this basis, the final score for the MT system is being computed. Figure 3 presents the evaluation of the noun phrases in figure 1. In this case, the hypothesis translation contains all the possible n-grams identified in the reference (6 variants for the 3-word phrase *le premier sommet*), so the instance receives the maximum score (6/6).

```
<Kybot id="kybot_a_n_de">
 <variables>
  <var name="X" type="term" pos="ART" />
  <var name="Y" type="term" pos="ADJ*" />
  <var name="Z" type="term" pos="NN*" />
 </variables>
 <relations>
  <root span="X" />
  <rel span="Y" pivot="X" direction="following" immediate="true" />
  <rel span="Z" pivot="Y" direction="following" immediate="true" />
 </relations>
 <events>
  <event eid="" target="$X/@tid" lemma="$X/@lemma" pos="$X/@pos"/>
  <role rid="" event="" target="$Y/@tid" lemma="$Y/@lemma" pos="$Y/@pos"
                                                        rtype="follows"/>
  <role rid="" event="" target="$Z/@tid" lemma="$Z/@lemma" pos="$Z/@pos"
                                                        rtype="follows"/>
 </events>
</Kybot>
```

Figure 2: A Kybot profile for a nominal group

4 MT Evaluation Method

Our extension with respect to the original version of the tool refers to the usage of alignments from the Alpine treebank. Previous papers on the topic [15, 9] had mentioned the limitation of automatically computed alignments and suggested methods to overcome the alignment noise, but none could compete the accuracy of a hand-aligned corpus. Therefore the strength of our approach consists in the integration of manually checked word alignments in the evaluation pipeline.

This required a special preprocessing step of extracting word alignments from the treebank and converting them to the format used by DELiC4MT. As an illustration, figure 4 depicts an aligned sentence pair from our treebank. Green lines indicate exact alignments and red lines represent fuzzy alignments. The corresponding alignments (from German to French) in the DELiC4MT format are:

```
0-0 1-2 2-3 3-4 4-5 5-6 5-7 6-8 6-9 6-10 6-11 7-12 8-14 8-15 ...
```

This means that the first word in the source sentence is aligned to the first one in the target sentence, taking into consideration that word numbering starts from 0.

The advantage of using a treebank over a word-aligned corpus is that the treebank contains other alignment types than “simple“ 1-1 word alignments. This often happens in German due to its frequent compounds, which are then represented as 1-n word alignments. For instance, the German compound *Montblanc-Abenteuer* (EN: Mont Blanc adventure) in figure 4 is aligned to the French noun phrase *aventure au Mont Blanc*. This correspondence is translated into the following word alignments: `6-8 6-9 6-10 6-11`.

```
Sen_id: 729 token_id: 6, 7, 8
Source tokens:  den, ersten, Gipfel
Alignments:  5-5, 6-6, 7-7,
Target equivalent ids: 5, 6, 7
Target sentence:
Après six heures j' atteignais le premier sommet, le Combin de la Tsessette.
Target equivalent tokens:  le premier sommet

Checking for n-gram matches for checkpoint instance: 319
Ref: le premier sommet
Hypo:
après six heures nous atteignons le premier sommet , le combin de la tsessette.
Number of 1-grams in reference: 3
# of matching 1-grams = 3
Number of 2-grams in reference: 2
# of matching 2-grams = 2
Number of 3-grams in reference: 1
Matched 3-gram: le premier sommet
# of matching 3-grams = 1
All n_gram matches :
le
...
le premier
premier sommet
le premier sommet

Total n_gram matches: 6
Total n_gram count in reference: 6
```

Figure 3: Sample output for a specific checkpoint instance

There are cases where a single word can correspond to a whole subtree in the other language. For example, the German adjective constituting the adjectival phrase *glücklich* (EN: happy) is paraphrased in French by the prepositional phrase *avec un sentiment de bonheur* (EN: with a feeling of happiness). We can thus use the phrase alignment in our treeebank and transform it into word alignments between the constituent words. In this way, our additional alignment level facilitates the extraction of n-m word alignments.

In order to demonstrate our claim, we have automatically computed the alignments for the 1000 sentences in the treebank with a well-established tool, GIZA++ [10]. Because the test set is relatively small for a statistical aligner to suggest accurate results, we have appended it to a considerably bigger corpus. For comparison purposes, we have chosen 200000 sentences from the Europarl corpus and, respectively, the same amount of sentences from an Alpine corpus. We have then used the generated alignments as input for the evaluation tool, along with the automatic translations generated by our SMT system (see section 5). Table 1 shows the results for several checkpoints for the language pair German-French.

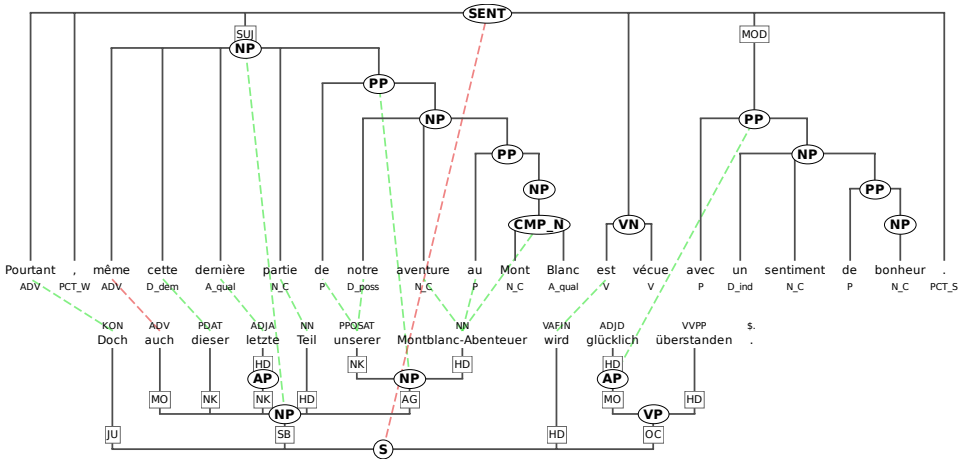

Figure 4: Aligned French-German tree pair from the Alpine treebank

The scores normally indicate the percentage of overlapping n-grams between the reference phrase (checkpoint instance) and the output produced by the MT system. However, in this context, the scores reported for the automatic alignments do not reflect the quality of the MT system. The evaluation module takes the same input in all three cases, except for the alignments, which are computed in different ways and generate different outcomes accordingly. Therefore the scores should be seen as estimates of the accuracy of the evaluation. The more precise the alignments, the more reliable the evaluation results.

We notice that the domain of the texts used for training GIZA++ does not influence significantly the accuracy, since the produced scores are similar (e.g. less than 2% difference between Europarl and the Alpine texts). However, when we compare the evaluation results with automatic alignments to the ones obtained with manual alignments, the latter ones are significantly better (up to 12% increase). This finding demonstrates the validity of our claim, namely that feeding manually proofed alignments from a parallel treebank to the evaluation pipeline generates more reliable results.

Checkpoint	**Alignment type**	**Final score**
Verb	GIZA++: Europarl	0.190 29
	GIZA++: Alpine	0.191 78
	Parallel Treebank	0.283 65
Det+Noun+Adj	GIZA++: Europarl	0.228 82
	GIZA++: Alpine	0.240 99
	Parallel Treebank	0.480 17

Table 1: Evaluation results for different alignments

Checkpoint	Instances	Google	PT	Our system
Noun	4790	0.313 72	0.351 29	**0.418 57**
Verb	953	0.211 94	**0.307 69**	0.283 65
Det_Adj_N	278	0.379 28	0.445 72	**0.480 17**
Dass_Pron_Verb	20	0.375 37	**0.383 56**	**0.383 56**
Verb_Pron_DetNoun	17	0.224 97	0.244 09	**0.409 45**
Weil_Pron_Verb	10	0.236 26	0.311 11	**0.577 78**
Pass	7	0.134 50	0	**1**

Table 2: Evaluation results for German-French

5 Evaluation Experiments

In this experiment, we compare our in-house SMT system with 2 other systems, Google Translate[8] and Personal Translator (PT)[9], in terms of handling specific linguistic checkpoints. Our SMT system was trained according to the instructions for building a baseline system at WMT 2011[10], with the difference that we use MGIZA++ [4] for computing the alignments. As training data we use Alpine texts from the Text+Berg corpus (approx. 200000 sentence pairs German-French).

The test corpus comprises 1000 sentence pairs from our Alpine treebank. For all systems, we use the manually-checked alignments extracted from the treebank. The comparison will be based on checkpoints which we considered particularly interesting for each translation direction, most of them PoS-based.

Table 2 contains the evaluation results for the language pair German-French. We have investigated the following checkpoints: nouns, finite verbs, noun phrases consisting of a determiner, an adjective and a noun (Det_Adj_N), subordinate clauses introduced by *dass* (EN: that) and *weil* (EN: because) and verb-subject-object collocations (Verb_Pron_DetNoun). Additionally, we have also considered the ambiguous word *Pass* (EN: passport, mountain pass, amble).

One notices that Personal Translator usually performs better than Google, probably because, being a rule-based system, it is aware of grammatical constructions and knows how to handle them properly. Its weaknesses are mostly related to the choice of words and unknown words, respectively. Since we are now looking at particular grammatical structures, it is likely for a rule-based system to analyze them adequately. Another evidence for this claim is the fact that Personal Translator outperforms all the other systems with respect to finite verbs, which pose difficulties in German (e. g. separable verbs).

Our in-house MT system performs in all cases better than its opponents because it has been trained with texts from the same domain. It thus gains strongly in vocabulary coverage. The most striking example is the German word *Pass* (EN:

[8] http://translate.google.com

[9] http://www.linguatec.net/products/tr/pt

[10] http://www.statmt.org/wmt11/baseline.html

Checkpoint	Instances	Google	PT	Our system
Noun_de_Noun	346	0.225 09	0.183 07	**0.415 81**
Poss_Noun	289	0.292 73	0.390 96	**0.469 55**
que_V	224	0.292 99	0.292 99	**0.340 76**
que_Pr_V	115	0.390 41	0.397 26	**0.445 21**
Ne_V_Pas	45	0.447 37	0.381 58	**0.526 32**
Verb_Pr_Vinf	25	0.158 73	0.301 59	**0.396 83**
V_Vinf_DetN	23	0.133 93	**0.294 64**	0.250 00

Table 3: Evaluation results for French-German

mountain pass), translated as *col* in the mountaineering domain, but as either *passe-port* (EN: passport) or *la passe* (EN: pass [the ball]) in other contexts. The analysis of the words with lemma *Pass* is reproduced in the last line of table 2. We notice that Personal Translator could not reproduce the correct meaning of the word in this context. Google succeeded getting one checkpoint instance right due to the collocation with a proper noun. As a result, it could correctly translate *Forcola-Pass* as *col Forcola*.

For the opposite translation direction, we have chosen linguistic structures particular for the source language (French): nominal phrases consisting of two nouns separated by the preposition *de* (such as in *journée d'été*) or of a possessive adjective and a noun (such as in *notre voisin*). Besides, we have selected relative clauses introduced by the word *que*, which can be interpreted as either a relative pronoun or a conjunction. Finally, we have considered verbal constructions: the negative form of verbs (e.g. *ne résisterait pas*), modal constructions involving nouns (e.g. *doit affronter le parcours*) and pronouns (e.g. *peut nous aider*), respectively. The results are presented in table 3.

The best-handled construction by all three systems is the negative verbal form, followed by the relative clause introduced by *que* and followed by a pronoun. If we remove the restriction that *que* is directly followed by a pronoun, the particle becomes ambiguous and this causes the drop of 10% between the scores in the third and the fourth line. Noun phrases are also handled well by our system, whereas complex verbal phrases raise challenges. The rule-based system Personal Translator gets the best score in the latter case due to its linguistic knowledge background, as we have noticed before.

5.1 Limitations of the System

As DELiC4MT evaluation uses string-based comparisons, it penalizes every small difference from the reference text. Consequently it will equally penalize a MT system for dropping a word as for misspelling a single letter from a word. This is particularly disadvantageous for SMT systems, which use no linguistic information (grammar, syntax or semantics). On the other hand, some of the limitations of

string-based comparisons can be easily overcome by considering not only word forms, but also lemmas or synsets. In the following, we outline some types of word form variations, which resulted in penalized errors:

- **Singular/Plural inconsistencies**: The tool distinguishes between singular and plural forms, although the word stem is the same. In the example below, the German sentence uses the singular form of the noun *Spur*, which is then translated in French as *la trace*. However, the reference translation suggests the plural form *les traces* in the given context, so the sentence pair is counted as a failed checkpoint, although the translation is fairly good.

  DE: *Im Abendrot bewundern wir die Spur unserer mutigen Vorgänger.*
  Automatic translation: *Au soleil couchant, nous pouvons admirer la trace de nos courageux prédécesseurs.*
  FR Reference: *Au coucher du soleil, nous admirons les traces de nos courageux prédécesseurs.*

- **Verbal tense inconsistencies**: If the MT system expresses the verbal phrase in a slightly different way, the tool will penalize the difference. For example, the finite verb *bewundern* in the previous example is translated as a modal construction: *pouvons admirer*. Since the French reference keeps the finite verbal construction, this checkpoint will also fail.

- **Compounds**: German compounds are a known challenge for SMT systems, because SMT systems do not possess a decomposition module. And when they finally get to be adequately translated, they fail to match the reference translation. For example, the compound noun *Montblanc-Expedition* is translated as *Montblanc expédition*. Since the reference translation was *expédition au Mont Blanc*, only a single n-gram matches, so the score for this checkpoint is very low (1/10).

- **Apostrophe words**: Another case which scores poorly in n-gram-based comparisons are word contractions, which are common in French, among others. This problem occurs mostly in conjunction with other MT errors, such as word choice. Suppose the evaluation tool has to compare the following instances of a pronoun-verb construction: the reference *j' aimerais bien* and the translation hypothesis *je voudrais*. The recall for this instance will be 0, since the system can not appreciate that the two pronouns (*je* and *j'*) are both variations of the first person singular in French. Moreover, the predicates also have different word-forms, although they convey the same meaning.

- **Synonyms**: As the previous example has showed, synonymy is not taken into consideration when comparing n-grams. Therefore, although phrases such as *un splendide après-midi* and *un magnifique après-midi* (EN: a wonderful afternoon) would perfectly match, they only get a score of 3/6.

6 Conclusions

In this paper we have described our experiments with the purpose of evaluating MT systems against a parallel treebank. We have demonstrated that we can improve the evaluation reliability by using manually checked alignments extracted from the treebank. In this way we could get an insight of the weaknesses of our MT system, by referring to problematic linguistic structures in the source language. In order to obtain a systematic classification of the problems, we should analyze the same structures in both languages.

We can conclude that this evaluation method does not offer a complete picture of the system's quality, especially because the output reduces to a number, as in the case of evaluation metrics. The advantage is that the score regards specific linguistic categories, rather than the overall performance of the system. In order to identify the source of the reported errors, a further manual analysis is needed.

The experiments have confirmed the advantage of using in-domain data for training SMT systems. Our system trained on a relatively small amount of in-domain training data (compared to the size of other corpora) outperforms systems not adapted to the domain. The better scores obtained by our SMT system in this evaluation scenario correlate with the BLEU scores reported in [12]. This finding proves the legitimacy of this evaluation approach, which is worthwhile to be extended, in order to obtain a more fine-grained analysis of the MT output.

References

[1] Anne Abeillé, Lionel Clément, and François Toussenel. Building a treebank for French. In *Treebanks : Building and Using Parsed Corpora*, pages 165–188. Springer, 2003.

[2] Wauter Bosma, Piek Vossen, German Rigau, Aitor Soroa, Maurizio Tesconi, Andrea Marchetti, Monica Monachini, and Carlo Aliprandi. KAF: a generic semantic annotation format. In *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon*, 2009.

[3] Elisabet Comelles, Jordi Atserias, Victoria Arranz, and Irene Castellón. VERTa: Linguistic features in MT evaluation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012. European Language Resources Association (ELRA).

[4] Qin Gao and Stephan Vogel. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, pages 49–57, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[5] Anne Göhring and Martin Volk. The Text+Berg corpus: An alpine French-German parallel resource. In *TALN 2011*, July 2011.

[6] Esther König and Wolfgang Lezius. The TIGER language - a description language for syntax graphs - Part 1: User's guidelines, 2002.

[7] Ding Liu and Daniel Gildea. Syntactic features for evaluation of machine translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005.

[8] Chi-kiu Lo and Dekai Wu. MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *Proceedings of ACL HLT 2011*, pages 220–229, 2011.

[9] Sudip Kumar Naskar, Antonio Toral, Federico Gaspari, and Andy Way. A framework for diagnostic evaluation of MT based on linguistic checkpoints. In *Proceedings of the 13th Machine Translation Summit*, pages 529–536, Xiamen, China, September 2011.

[10] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March 2003.

[11] Maja Popović and Hermann Ney. Word error rates: decomposition over POS classes and applications for error analysis. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 48–55, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[12] Rico Sennrich. Combining multi-engine machine translation and online learning through dynamic phrase tables. In *EAMT-2011: the 15th Annual Conference of the European Association for Machine Translation*, May 2011.

[13] Rico Sennrich and Martin Volk. MT-based sentence alignment for OCR-generated parallel texts. In *The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010)*, 2010.

[14] Martin Volk, Torsten Marek, and Yvonne Samuelsson. Building and querying parallel treebanks. *Translation: Computation, Corpora, Cognition (Special Issue on Parallel Corpora: Annotation, Exploitation and Evaluation)*, 1(1):7–28, 2011.

[15] Ming Zhou, Bo Wang, Shujie Liu, Mu Li, Dongdong Zhang, and Tiejun Zhao. Diagnostic evaluation of machine translation systems using automatically constructed linguistic check-points. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 1121–1128, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.